A Comparison of Three Methods for Automatic Text Categorization for Geriatrics

Josh Denny, MD, James Powers, MD, Anderson Spickard, III, MD, MS, Randolph Miller, MD
Vanderbilt University School of Medicine, Nashville, TN

Introduction
Relevant to clinical education and practice, text categorization is a common information retrieval task. Previously, we developed KnowledgeMap (KM), which enables concept-based classification of medical school curricular document contents using relationships defined in the UMLS Metathesaurus and the UMLS Semantic Network. Users can retrieve curricular content using KM via both specific-phrase queries and broad-term queries (e.g., “where is genetics taught?”). Using as the “gold standard” test documents categorized by clinical experts, KM achieves sensitivities of 0.78 – 1.00 and specificities of 0.85 – 0.96 for broad queries; area under receiver operator characteristic curves varies from 0.89 -0.98.

Broad queries require pre-existing, well-defined expansion hierarchies for the broad search terms (e.g., genetics above), in order to identify documents relevant to various specific terms derived from the broad query term.

With the goals of delivering of just-in-time educational content during geriatric patient encounters and tracking curriculum development over time, we attempted to classify documents for the broad concept of “geriatrics” by comparing three approaches. The problem is made difficult by the lack of a well-defined expansion hierarchy for terms related to the concept “geriatrics” in the UMLS.

Methods
Our approaches to retrieving curricular documents relevant to “geriatrics” included: first, existing KM, which expands the UMLS concepts of “aging”, “elderly”, and “geriatrics” as parent concepts at two levels of expansion, yielding 251 and 4,509 related concepts, respectively; second, defining “geriatrics” through use of concepts extracted (via KM) from an external, geriatrics-rich resource, the American Geriatric Society’s Geriatric Review Syllabus (GRS); and third, a support vector machine (SVM) based approach.

Our gold standard document set came from the 2006 Vanderbilt medical curriculum, with balanced representation of preclinical and clinical documents. Two authors (JP and RM) each scored about 200 documents from the set of 337, with fifty documents overlapping between them. The SVM model operated on normalized words and UMLS concepts within the expert-ranked gold standard documents weighted by frequency. We used a polynomial SVM with a 5-fold cross-validation to optimize parameters across degree and cost. We ranked concept terms extracted from GRS using a term frequency-inverse document frequency (TF-IDF) weighting scheme that contrasted the GRS with a non-geriatric curriculum using relative frequency, breadth, and the semantic types of the concepts. We then scored each of the 337 documents with the three algorithms, and calculated the area under the receiver operator characteristic curves (AUC), comparing them with the $\chi^2$ statistic.

Results
Of the 337 curricular documents scored by the reviewers, 285 (85%) were rated low-relevance, 45 (13%) rated moderate-relevance, and 7 (2.1%) rated as high-relevance. Interrater agreement (kappa) between the reviewers was 0.55. KM performed similar to or better than other algorithms (see Table).

Conclusions
Existing KM algorithms and SVMs classified documents similarly; however, performance was not as high as in prior experimentation on other topics. This likely reflects terminology deficiencies; important geriatrics concepts such as “incontinence” and “falls” were notably absent in the automated KM expansions. Semantic type filtering substantially improved performance of the GRS curriculum by devaluing semantic types with lower predictive values (e.g., “intellectual product” or “functional concepts”). Augmenting KM expansions with selected concepts from a source such as the GRS curriculum may improve results.

This work was supported in part by a Geriatrics Education Center grant from the D.W. Reynolds Foundation.

References